intel

ARTICLE
REPRINT

AR-260

March 1983

E-PROMs Graduate To 256-K Density With Scaled N-Channel Process

M. Van Buskirk, M. Holler, G. Korsh,
B. Lee, S. Lee, D. Tang, G. Teng,
S. Fouts, P. Dang, and W. Fisher

MARCH 1983
ORDER NUMBER: 210988-001

Technical articles

E-PROMs graduate to 256-K density with scaled n-channel process

With 32-K bytes per chip, erasable programmable read-only memory can carry application software for business and personal computers

by M. Van Buskirk, M. Holler, G. Korsh, B. Lee, S. Lee,
D. Tang, G. Teng, S. Fouts, P. Dang, and W. Fisher, *Intel Corp., Santa Clara, Calif.*

☐ Since the introduction of the 2-K 1702 in 1971, the erasable programmable read-only memory has shaken off its reputation as a mere prototyping tool and emerged as a major commodity, worth some $240 million last year in the U. S. alone. Central to that growth, a heady pace of process and circuit innovations has doubled E-PROM densities every one to two years.

The advent of the 256-K chip—exemplified by the 27256 from Intel—signals the crossing of key technical hurdles and with it an open path to accelerated development of megabit and larger arrays. Though clearly the child of earlier generations of E-PROMs, the part has been thoroughly scaled down to achieve thinner oxides and minimum features of 1 micrometer. New sensing and decoding circuitry are incorporated as well.

Along the path to even larger arrays, the price per bit of E-PROMs will continue to drop, closing in on that of ROM, the least expensive semiconductor storage. Indeed, the 256-K ROM has only a year's jump on the 27256. The cost of E-PROM is projected at just 50% more than that of ROM in 1985: 6 versus 4 millicents per bit.

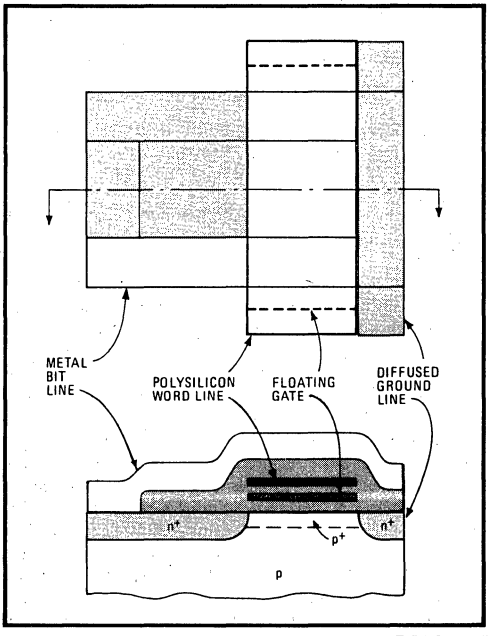
Achieving 256-K density in an E-PROM calls for a host of advances working in concert, dramatic scaling down of device dimensions being only the most obvious one. Processing changes also accompany the smaller geometries and thinner layers. New designs for the decoding and sensing circuitry cope with the worsening problem of statistical variations in device parameters among the more than quarter million bits in the array.

The flexibility of alterable nonvolatile storage coupled with the economy of high chip density promises intriguing possibilities for the architect of microsystems, particularly portable computers for the mass market. As the table on page 93 suggests, significant system and application software fits in a very few 256-K arrays. For example, two or three chips can carry a Pascal compiler and a sophisticated word-processing program.

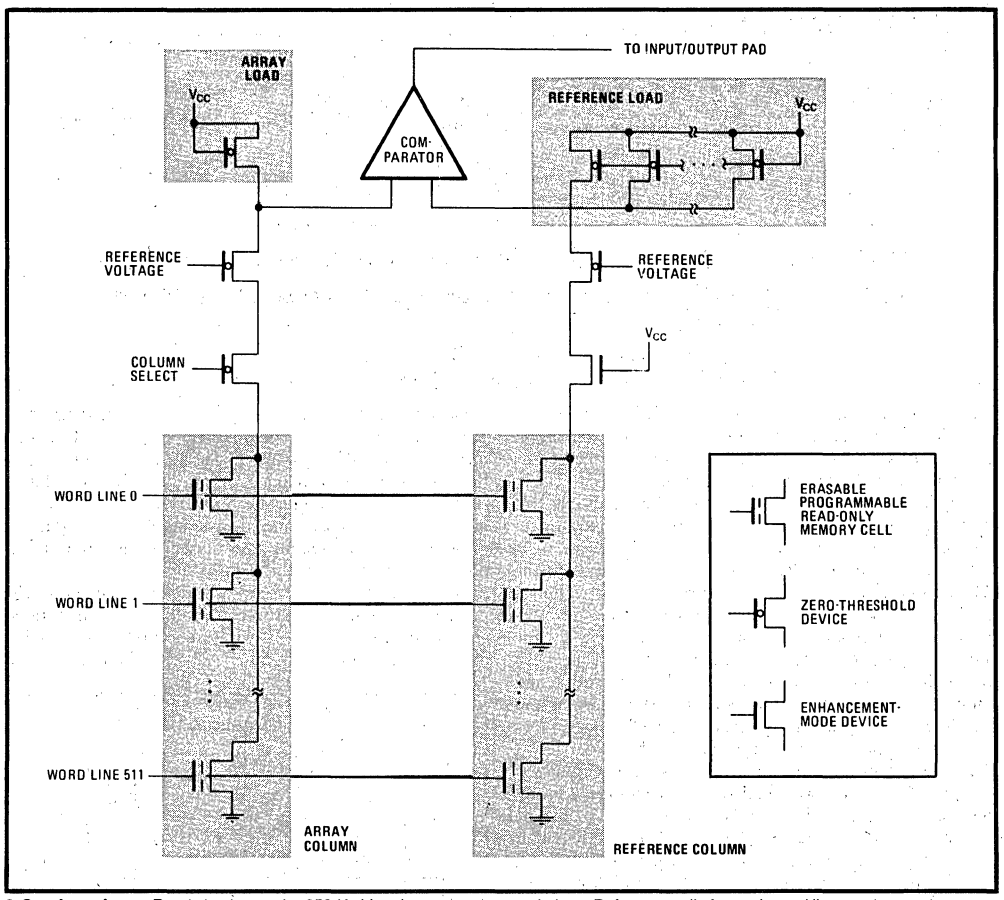
To the user of a machine incorporating such firmware, the friendliness and convenience of fast, 200-nanosecond memory accessed with the push of a button contrasts sharply with the fuss and delay of floppy disks. Further, for the end user and equipment maker alike, E-PROM continues to offer compelling advantages over ROM, advantages that will come at an ever lower premium as E-PROMs succeed in catching up with ROMs in density.

Speeding a product to market, for example, often means last-minute software changes, a costly and impractical requirement with mask-programmable chips, but a simple matter for E-PROMs. Similarly, the E-PROM can hold down the costs and delays involved in updating programs or in correcting those last few bugs that somehow made it out into the field.

Scaling chip geometries down in size has been the primary instrument of progress in all types of integrated circuits in the last 6 to 10 years. The first major scaling down of E-PROMs occurred in 1980 with the introduction



1. Scaled. Two-micrometer design rules squeeze the E-PROM cell down to 6 by 6 μm. The active channel area, beneath the floating polysilicon gate, is just 1 by 1.2 μm. The n^+ regions are 0.5 μm deep.



2. Sensing scheme. Read circuitry on the 256-K chip tolerates process variations. Reference cells for each word line monitor read currents; identical geometry for the array and reference load transistors ensures a constant current ratio for sensing.

of the 2764 64-K part with a 159-square-micrometer cell and H-MOS (high-performance MOS) peripheral circuits. The major technology advance prior to that was the conversion in 1977 to n-channel MOS technology with depletion-mode devices for the 16-K 2716, a move that made operation possible from a single 5-volt power supply.

A completely redesigned process and second major scaling in 1982 have resulted in another leap in density and performance for E-PROMs, producing a 256-K array with a 6-by-6-μm cell, for a chip size of just 28,500 square mils. The floating-gate storage transistor, with a channel effectively 1 μm long by 1.2 μm wide, employs a first oxide 325 angstroms thick and a second of 400 Å (Fig. 1). The peripheral circuitry, with a 0.5-picojoule speed-power product, is equivalent to chips built in H-MOS II, a second-generation high-performance MOS technology that uses 2-m channel lengths and 400-Å gate oxides for minimum gate delays of 0.4 ns.

The practical constraint of maintaining the same 5-V power supply used in previous technologies while scaling transistor sizes stresses the materials constituting the devices: both substrate and gate dielectrics are exposed to much higher electric fields. The fabrication process must build a margin of safety into the chip to prevent unwanted effects like time-dependent oxide failure, pn-junction breakdown, and parasitic MOS-transistor action between adjacent diffused regions.

In E-PROMs, the high voltage required for programming the cells further aggravates the situation. In fact, in the storage transistor, the oxide surrounding the floating gate must be flawless, or else electrons leak off, losing the stored data. For that reason, the programming voltage in the 27256 is scaled down to 12.5v (see "Scaling down the E-PROM cell," opposite). With that scaling, the 400-A oxides in the transistors of the peripheral circuitry that route the programming voltage to the array experience an electric field of 3.25 megavolts per centimeter, rough-

Scaling down the E-PROM cell

A two-step scaling-down process starting from the 64-K erasable programmable read-only memory (the 2764) has resulted in the 256-K chip designated the 27256 by Intel. The original 64-K E-PROM represented a modest scaling of the 16-K, using positive photoresists and projection-printer lithography. The first step beyond the 64-K chip called for wafer-stepping lithography for levels such as the first polysilicon and contact openings, where significant area could be saved without redesigning the devices or changing the other process steps. Those moves substantially reduced the size of the 64-K chip and ushered in a 128-K version.

At the same time, however, a program was under way to scale down all the design rules and use the stepper lithography to full advantage. That complete redesign of the process not only produced the 256-K array, but suggests that, from a device design perspective, 512-K and 1-megabit arrays will be feasible in the next two to three years. However, substantially better control over dimensions will have to accompany that continued scaling. Several variations were made upon the so-called classical scaling theory and can be understood with reference to the operation of the E-PROM cell.

The E-PROM cell is a simple modification of a conventional n-channel MOS enhancement-mode transistor whose drain connects to a bit line and whose source is grounded. The transistor's gate floats and is controlled by capacitive coupling to a polysilicon word line overlying the gate. Current conducted through the transistor is read as a logical 1, and the absence of current as a 0.

The cell is programmed to preserve a nonconducting state by applying high voltages to the word and bit lines simultaneously. Under those conditions, hot electrons are injected from the channel to the floating gate, charging it to a negative voltage. With the gate charged by the trapped electrons, the word line cannot couple enough voltage to it to turn on the transistor during a subsequent read operation. Exposing the cell to ultraviolet light elevates the trapped electrons' energy to the level of the conduction band of the surrounding oxide, and their mutual repulsion then causes them to flow off the gate.

The cell for the 256-K chip was derived from that of its 64-K predecessor using a scaling factor of 2. As shown in the table, the constant-field scaling theory was followed almost exactly for the device dimensions, oxide thicknesses, and programming voltage. Significant departures from the formulas were made in the read voltage, which was not scaled at all, and the channel-doping concentration, which was more than doubled (in fact almost quintupled). The doping was boosted this much to increase the electric fields in the channel and thereby improve the programming efficiency. Such an increase was feasible because the read voltage was not scaled: the threshold voltage is set as a fraction of the read voltage to ensure the optimal logic threshold, and it therefore could also remain unchanged.

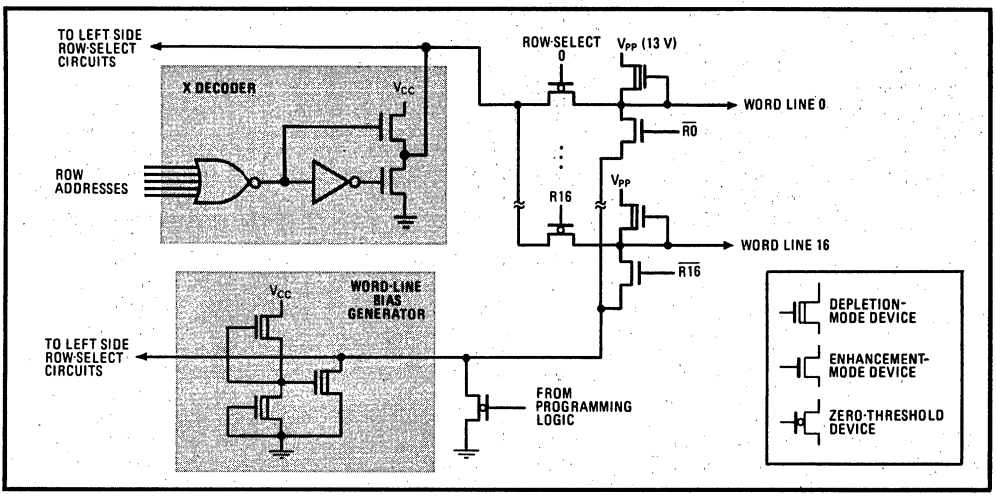
In the scaling theory, the read voltage would be scaled to prevent an increase in the electric field—and hence the stress—across the oxides. In an E-PROM, however, special treatment is required for the oxides, which see a worst-case field due to the high programming voltage. Thus, with the programming voltage scaled and oxides with the requisite integrity fabricated, the read voltage presents no problem. In fact, leaving the read voltage at 5 volts instead of scaling it to 2.5 V substantially increases the read current and speeds up access to the array.

Because of the high channel fields caused by the high dopant concentration, a 5-V drain voltage could suffice to inject hot electrons to the gate during a read, an effect appropriately known as parasitic programming. Therefore, the drain voltage was scaled down well below 5 V.

The devices in the peripheral circuitry were scaled from an effective length of 3.4 micrometers to 1.7 μm, and their oxides were thinned from 650 angstroms to 400 Å, which is equivalent to H-MOS II transistors. Those two changes increased the current of a minimum-length transistor by a factor of only 2.3: the increase in channel doping of the enhancement-mode devices degrades the electron mobility, limiting the current increase. Because the supply voltage was not scaled, the speed-power product of the technology is greater than it otherwise might have been, decreasing only by half to 0.5 picojoule.

SCALING DATA FOR THE 256-K ERASABLE PROGRAMMABLE READ-ONLY MEMORY

Parameter	64-K chip	Constant-field scaling formula	Theoretical 256-K chip	Actual scaling formula	Actual 256-K chip
Cell area	159 μm^2	K^{-2}	40 μm^2	$\approx K^{-2}$	36 μm^2
Floating-gate oxide	725 Å	K^{-1}	363 Å	$\approx K^{-1}$	325 Å
Control-gate oxide	900 Å	K^{-1}	450 Å	K^{-1}	450 Å
Channel doping	2×10^{16} cm^{-3}	K	4×10^{16} cm^{-3}	$K^{1.7}$	9×10^{16} cm^{-3}
Threshold voltage	1.6 V	K^{-1}	0.8 V	1	1.6 V
Cell current	55 μA	K^{-1}	28 μA	$K^{+0.6}$	80 μA
Read voltage	5 V	K^{-1}	2.5 V	1	5 V
Program voltage	21 V	K^{-1}	11 V	$\sim K^{-0.7}$	12.5 V



3. Biased decoder. With a 0.5-volt bias applied to deselected word lines during a read operation, depletion-mode pass transistors can be used in the decoder. They increase the selected word line's voltage, speeding access and extending the allowed power-supply range.

ly equal to the electric field found in previous parts.

The thin oxide between the substrate and floating gate illustrates the engineering that went into each of the process steps. Early in the development, that 325-Å-thick oxide showed defect densities of around 100/cm². (The oxide integrity is judged with a sensitive measurement of the dc current flowing through a large-area capacitor biased to 15 V.) Based on the total active area, A, of the channels of the 256-K chip's array transistors, a simple model for the yield, such as e^{-AD}, predicts the loss due to that defect density, D, at 54%, an unacceptably high figure for a single process step.

With careful adjustment of process parameters, the thin-oxide defect density was reduced below 5/cm², raising the yield of that step to an estimated 96%. Similarly, each new step could be fine-tuned independently of any other procedures, simplifying the process debugging. Despite the many new procedures, the overall structure and the sequence of steps do not depart radically from previous E-PROMs, a help in the retraining of manufacturing personnel.

In particular, the shared drain contact, common source diffusion, and self-aligned floating polysilicon gate of the E-PROM cell have not changed conceptually since the 2716 of five years ago. What has changed dramatically is the minimum feature size: the evolution from projection printing and wet etching to wafer-stepping lithography and anisotropic plasma etching on critical levels has shrunk the 27256's cell to the size of a contact hole on the old 2716.

Most elements of the E-PROM process are affected by that radical shrink. The field oxide that isolates adjacent diffusions is thinned to 0.6 μm, a move that cuts the length of the bird's beak in half. (The bird's beak, or transition region, between the oxide and the active device is wasted area.) As a result, the minimum spacing on the mask between adjacent field oxide regions is only 2 μm. The thinner oxide is possible thanks to the lower programming voltage, which reduces the chance of parasitic transistor action beneath the oxide.

The dose of boron implanted in the transistor channels to set their threshold voltage is increased, compensating for the effects of the shorter, narrower channels. In turn, the gate oxides are thinned to boost the transconductance of the more highly doped channels. The arsenic for the source and drain regions is implanted 0.5 μm or shallower to control the channel length and forestall punchthrough. To prevent the metalization from penetrating those shallow junctions, an aluminum-silicon alloy substitutes for the usual aluminum.

Although process development goes far to ensure a reliable part with good margins, new circuits were also required for the E-PROM's jump to 256-K density. For one, some statistical variation in cell characteristics is inevitable, and the more bits per chip, the greater the likelihood that one or more bits will fall outside the acceptable limits. Thus, circuits that compensate for these process-induced variations in effect raise yield. Sensing circuitry offers one example.

The 27256 uses two entire columns of reference cells, associating two cells with each of the 512 word lines rather than one with each of the eight comparators in the output section (Fig. 2). The tolerance to variations in read currents quintuples that in the 2764. Undoubtedly, future E-PROMs will extend this concept again, incorporating multiple reference columns.

Before the 2764, E-PROMs employed single-ended sense amplifiers, which are comparators that switch when the input current exceeds some threshold, called the trip current, which is not a function of any cell parameters. To make the trip current dependent on the read current itself, the 2764 used a differential comparator, which compares the read current with a fixed fraction of the current from a reference cell—a replica of

the array cells. If a cell's read current is lower than expected, then the reference cell's current will probably be low as well, adjusting the trip current accordingly. Such a circuit not only continues to operate under process variations, but also allows the design of a comparator optimized for speed.

Unfortunately, carrying that scheme even further on the 27256 and increasing the number of reference cells naturally leads to a wider distribution of their currents, and the conventional differential comparator could introduce an error because of that variation. To compensate for that error, a constant-current–ratio differential comparator was introduced. In it, the lower impedance reference load is obtained by connecting several transistors in parallel, each of which is an identical copy of the array load device (see Fig. 2 again). The parallel connection then gives the correct impedance without changing the transistor operating point, permitting correct operation over a much wider range of read current.

Extending a scheme

In the usual comparator, the dimensions of the load device on the reference cell or column are adjusted to give a lower impedance than that of the load device on the array column. The voltage drops across the different impedances are the differential input to the comparator. However, the different load-device dimensions change the devices' threshold voltages, putting the reference load at a different operating point. That difference means that the ratio of the array current to the reference current changes with variations in the reference current.

Another innovation, in the row-decoder circuits, helps speed access time and extend the power-supply variations the chip will tolerate. The scheme devised for the 27256 uses a negative-threshold pass transistor; however, it also incorporates a bias-voltage generator to hold the deselected word lines at about 0.5 V, rather than 0 V, during a read operation (Fig. 3). The bias current for the generator is supplied by pull-up transistors connected to the programming voltage supply. The 0.5-V shift suffices to reduce the leakage current without turning on deselected cells during a read operation.

However, during programming, when the high voltage applied to selected columns can couple to the floating gates, the 0.5-V bias would be enough to partially turn on deselected cells. Thus, a shunt transistor is included in the bias generator to pull the deselected word lines all the way to ground during programming.

In a conventional decoder, if a positive-threshold device is used, the word-line voltage rises very slowly above the level of one threshold below the supply voltage, slowing access to the cells. It also ups the minimum power-supply voltage by about 1 V. (The minimum supply voltage is the cell's threshold voltage, plus that voltage required to generate a detectable current difference between programmed and erased cells, typically about 3 V.)

Using a negative-threshold device overcomes both those effects, but introduces its own problems if not accompanied by a bias-voltage generator. With a depletion-mode pass transistor, the deselected devices still pass a significant leakage current. That current pulls down the decoder output below the supply voltage so that the selected word line still receives a reduced voltage.

TYPICAL PROGRAM SIZES FOR 256-K MEMORIES

Program	Bytes	Number of 256-K erasable programmable read-only memories
Basic interpreter	20-K to 32-K	5/8 to 1
Pascal compiler	32-K to 40-K	1 to 1 1/4
Asteroids	8-K	1/4
Screen-oriented editor	10-K	1/3
Word processor	32-K to 48-K	1 to 1 1/2
Spelling dictionary	12-K	3/8
Relational data base	32-K to 128-K	1 to 4

The 27256 also illustrates the utility of two circuit features trademarked as the intelligent Identifier and the intelligent Programming Algorithm. A persistent problem for the user of scaled-down E-PROM technology is the changing programming requirements. Each new generation forces customers to convert to lower voltages and altered algorithms. The identifier, an on-chip, unalterable, 2-byte code, specifies the chip's manufacturer, programming algorithm, voltage, and pulse width.

Saving progamming time

With ever denser E-PROM arrays, the programming algorithm proves its worth in saving programming time. This adaptive, closed-loop algorithm varies the width and the number of program pulses to reach an adequate margin in the minimum time, typically 4 milliseconds per byte for the 27256.

The old programming algorithm, developed for the 2716, used a fixed 50-ms pulse width. That width is determined by the worst-case programming time, that is, the longest time any bit in the array will need for complete programming. As with other device characteristics, process variations lead to a distribution of programming times; the fixed-width pulse must be long enough to program the slowest bit in the array. At the 256-K level, chances are good that at least one bit would need a very long programming pulse.

The closed-loop programming algorithm requires a method of determining the programmed cell "margin," or retention characteristics. Margin has generally been represented as the maximum supply voltage at which both programmed and erased cells can be read. A programmed cell is one in which the threshold voltage has been forced to a higher voltage. Increasing the supply voltage boosts the voltage applied to a cell's gate; eventually, the higher threshold is reached, turning on a cell that is intended to remain off.

Thus, the strategy in the programming algorithm is to verify that a cell is programmed at an elevated supply voltage of 6 V. The procedure begins by applying a 1-ms pulse to the first byte. At the end of the pulse, the chip tries to read the programmed cells. If they are programmed, an additional 3-ms pulse is applied, boosting the margin about 1.5 V. If the cell is not programmed, 1-ms pulses continue to be applied until programming is verified. When it is, a pulse three times as long as the sum of the 1-ms pulses already applied adds the required margin. To program a 27256 takes about 3 minutes. □